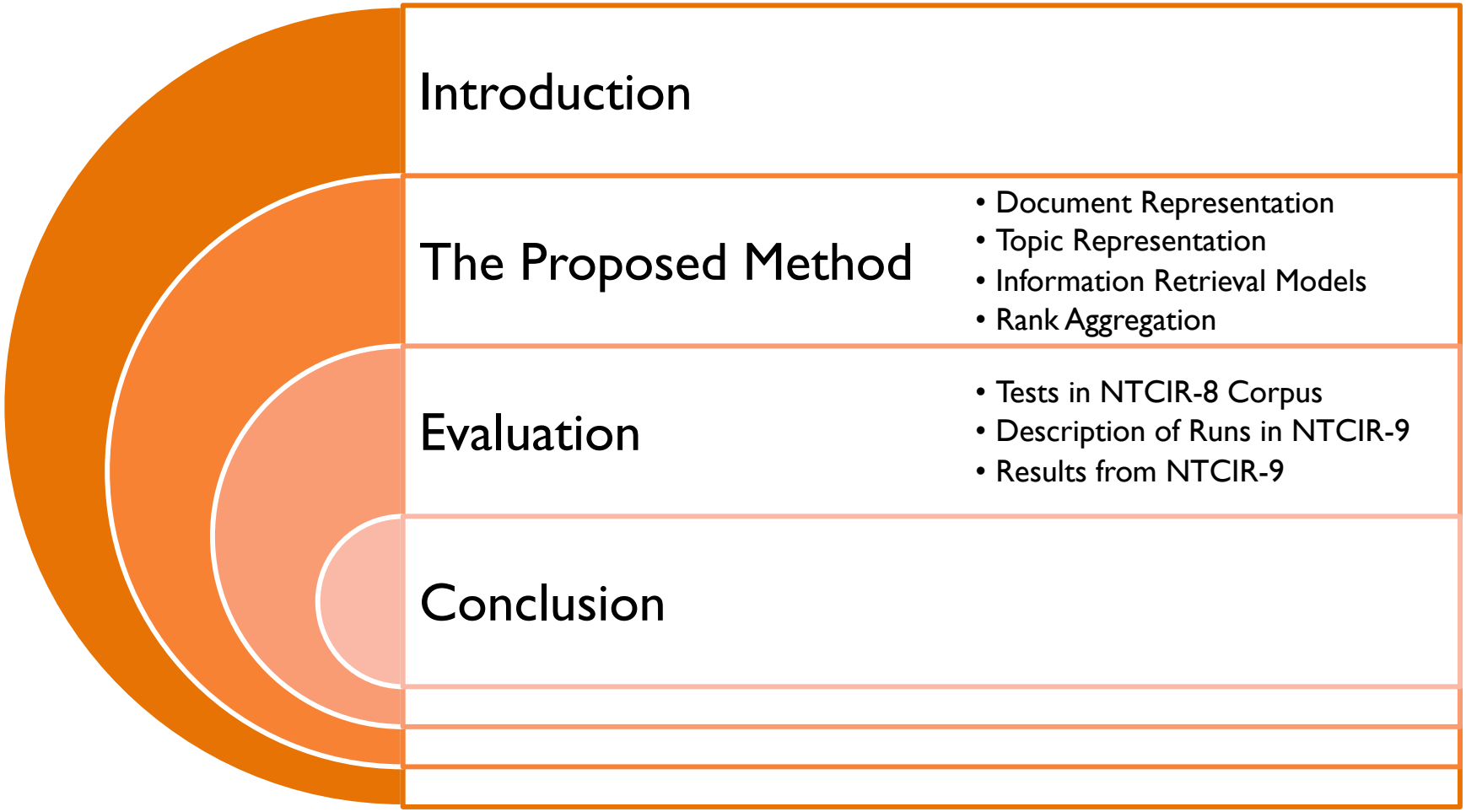
NTCIR-9 GeoTime:

Geo-temporal Information Retrieval Based on Semantic Role Labeling and Rank Aggregation

Yoonjae Jeong, Gwan Jang, Kyung-min Kim and Sung-Hyon Myaeng
Korea Advanced Institute of Science and Technology (KAIST)

Contents


Introduction
The Proposed Method
• Document Representation
• Topic Representation
• Information Retrieval Models
• Rank Aggregation
Evaluation
• Tests in NTCIR-8 Corpus
• Description of Runs in NTCIR-9
• Results from NTCIR-9
Conclusion

Introduction (1/2)

- NTCIR-9 GeoTime Task
  - It is about geographic and temporal search in news articles.
  - We participated in the English sub-task only.

- Topics asks for geographic- and temporal-based information.
  - A topic asks for information on where and when a particular event occurred or what event happened at a specific time and location.
  - "where, when, and what did <*entities*> <*action*>?"

- The proposed geo-temporal information retrieval model
  - Our basic idea is to add locational and temporal aspects to terms in a document using Semantic Role Labeling (SRL).

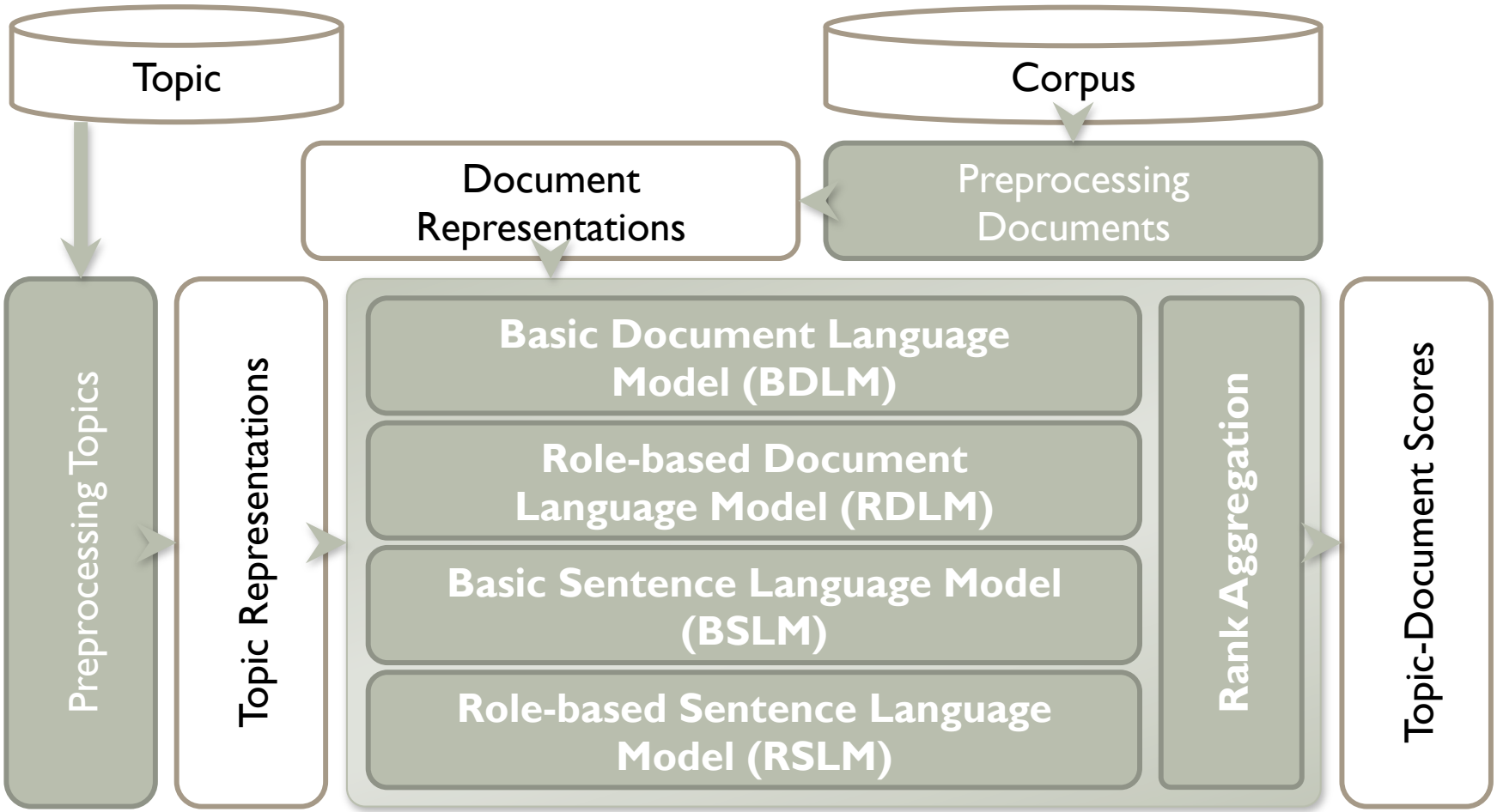
Introduction (2/2)

- Semantic Roles
  - A semantic role is the underlying relationship that a participant (linguistic constituent) has with the main verb in a clause.
  - In the form, the elements of topic are reflected in following semantic roles.

Topic	Semantic Role
where	AM-LOC
when	AM-TMP
<*entity*>	A0-5 (e.g., Agent, Patient, …)
<*action*>	verb

The Proposed Method

Overview of the proposed geo-temporal information retrieval



The Proposed Method

Document Representations

- Documents are represented as sets of words for semantic roles.
  - An example of SRL for a document

> **[A-1** Astrid Lindgren, the Swedish writer whose rollicking, anarchic books about Pippi Longstocking horrified a generation of parents and captivated millions of children around the globe], died in her sleep **[AM-TMP** Monday] **[AM-LOC** at her home in Stockholm, Sweden.]

Attribute	Description
T_V	A set of verb in document e.g., *die*
T_A	A set of terms with numbered argument roles (A0-5) in document e.g., *Astrid*, *Lindgren*, … , *children*, *globe*
$T_{AM\text{-}LOC}$	A set of terms with location (AM-LOC) roles in document e.g., *home*, *Stockholm*, *Sweden*
$T_{AM\text{-}TMP}$	A set of terms with temporal role (AM-TMP) in document e.g., *Monday*

The Proposed Method

Topic Representations (1/2)

Attribute		Description
Question types	Q-LOC	Whether a question is about location or not? e.g., *When and <u>where</u> did Astrid Lindgren die?*
	Q-TMP	Whether a question is about time or not? e.g., *<u>When</u> and where did Hurricane Katrina make landfall in the United States?*
	Q-AGT	Whether is a question about agent or not? e.g., *<u>What Portuguese colony</u> was transferred to China and when?*
	Q-MSC	The others e.g., *How old was Max Schmeling when he died, and where did he die?*
Set of vocabularies	V_V	A set of vocabularies in verb role in topic
	V_A	A set of vocabularies in numbered argument (A0-5) roles in topic
	$V_{AM\text{-}LOC}$	A set of vocabularies in locational role (AM-LOC) in topic
	$V_{AM\text{-}TMP}$	A set of vocabularies in temporal role (AM-TMP) in topic

The Proposed Method

Topic Representations (2/2)

- To determine the question types, we devised some heuristic rules based on syntactic parser results.
  - A parse-tree example for question type identification

```
Topic: When and where did Astrid Lindgren die?
Parsing Tree:
(ROOT
  (SBARQ
    (WHADVP (WRB When)
      (CC and)
      (WRB where))
    (SQ (VBD did)
      (NP (NNP Astrid) (NNP Lindgren))
      (VP (VB die)))
    (. ?)))
```

Information Retrieval Models (1/4)

- Basic Document Language Model (BDLM)
  - Basic Language Model for a topic *q* and document *d*.

$$P_{BDLM}(d|q) = P(q|d) \times \frac{P(d)}{P(q)} \approx \prod_{t \in q} P(t|d) \quad (1)$$

$$P(t|d) = \frac{tf_{t,d} + \mu \times P(t|D)}{|d| + \mu} \quad (2)$$

d : a document
q : a topic
t : a term in *q*
$tf_{t,d}$: the term frequency of term *t* in document *d*
D : the set of all document in the corpus
μ : the smoothing parameter (=2500)

Information Retrieval Models (2/4)

- Role-based Document Language Model (RDLM)
  - Based on the document representation, we built the *RBLM* where *R* represents the semantic roles in the document *d*.

$$P_{RDLM}(d|q) = \prod_{r \in R} \left(P(q_r|d_r) + \alpha \right)$$

$$\begin{cases} \alpha = 1, & \text{if Q-LOC is } \textit{true}, r\text{=AM-LOC and } |d_r| > 0. \\ \alpha = 1, & \text{if Q-TMP is } \textit{true}, r\text{=AM-TMP and } |d_r| > 0. \\ \alpha = 0, & \text{otherwise} \end{cases} \quad (3)$$

R: the semantic roles in q
q_r : the set of terms given the role r in q
d_r : the set of terms given the role r in d

Information Retrieval Models (3/4)

- Basic Sentence Language Model (BSLM)
  - Sometimes the relevant information related to a topic is fully contained in one sentence in document.

$$P_{BSLM}(d|q) = \max_{s \in S} P(s|q) = \max_{s \in S} \prod_{t \in q} P(t|s) \quad (4)$$

S : a set of sentence in d
s : a sentence in S

Information Retrieval Models (4/4)

- Role-based Sentence Language Model (RSLM)
  - RSLM adds semantic roles to BSLM in the same way RDLM was constructed out of BDLM.

$$P_{RSLM}(d|q) = \max_{s \in S} \prod_{r \in R} \left(P(q_r|s_r) + \alpha \right)$$

$$\begin{cases} \alpha = 1, & \text{if Q-LOC is } \textit{true}, r\text{=AM-LOC and } |s_r| > 0. \\ \alpha = 1, & \text{if Q-TMP is } \textit{true}, r\text{=AM-TMP and } |s_r| > 0. \\ \alpha = 0, & \text{otherwise} \end{cases} \quad (5)$$

R : the semantic roles in q
q_r : the set of terms given the role r in q
s_r : the set of terms given the role r in s

The Proposed Method

Rank Aggregation (1/4)

- Relevant documents and their ranks and normalized scores of each model for Topic GeoTime-0025 in NTCIR-8 corpus

Relevant Document	Rank & Normalized Score			
	BDLM	RDLM	BSLM	RSLM
NYT_ENG_20041226.0096	24 (8.95E-01)	201 (1.54E+00)	251 (3.72E-02)	**4 (2.82E+02)**
NYT_ENG_20041229.0208	167 (1.07E-04)	322 (1.07E+00)	133 (9.18E-02)	**3 (2.86E+02)**
NYT_ENG_20041230.0186	**3 (2.20E+02)**	18 (4.46E+01)	298 (3.09E-03)	102 (7.74E-01)
NYT_ENG_20041230.0204	**8 (8.47E+01)**	26 (3.26E+01)	302 (3.09E-03)	98 (7.75E-01)
NYT_ENG_20041230.0245	**4 (1.21E+02)**	17 (4.46E+01)	299 (3.09E-03)	101 (7.74E-01)
NYT_ENG_20041230.0256	**6 (1.20E+02)**	25 (3.49E+01)	303 (3.09E-03)	100 (7.74E-01)
NYT_ENG_20041231.0009	**2 (2.20E+02)**	19 (4.46E+01)	300 (3.09E-03)	99 (7.75E-01)
NYT_ENG_20050328.0205	36 (4.30E-03)	349 (1.07E+00)	88 (9.18E-02)	162 (7.72E-01)

Rank Aggregation (2/4)

- The proposed IR models have difference characteristics.
- It can be combined to handle various retrieval cases by devising a rank aggregation method, which is based on Dwork, et al. (2001)'s Markov Chain based approaches (MC1, MC2, MC3, & MC4).

- We adopted MC2 because it is arguably the most representative of minority viewpoints of sufficient statistical significance; it protects specialist views.

The Proposed Method

Rank Aggregation (3/4)

- The transition matrix of the k^{th} rank list τ_k, T_k is

$$T_k \triangleq \left(t_{ij}^{(k)}\right)_{n\times n} \tag{6}$$

$$t_{ij}^{(k)} = \begin{cases} \dfrac{1}{\#\left\{j \,\middle|\, j >_{\tau_k} i \text{ or } i = i\right\}}, & j >_{\tau_k} i \text{ or } j = i \\ 0, \quad \text{otherwise} \end{cases} \tag{7}$$

$j >_{t_k} i$: document j is ranked higher than document i in ranking list τ_k.

- The final transition matrix T is

$$T = \frac{1}{l}\sum_{k=1}^{l} T_k \tag{8}$$

l : the number of ranked lists

Rank Aggregation (4/4)

- The effective ranks for aggregations of RDLM, BSLM, and RSLM are a small number of top ones.
- We applied the threshold θ to the elements of transition matrix.

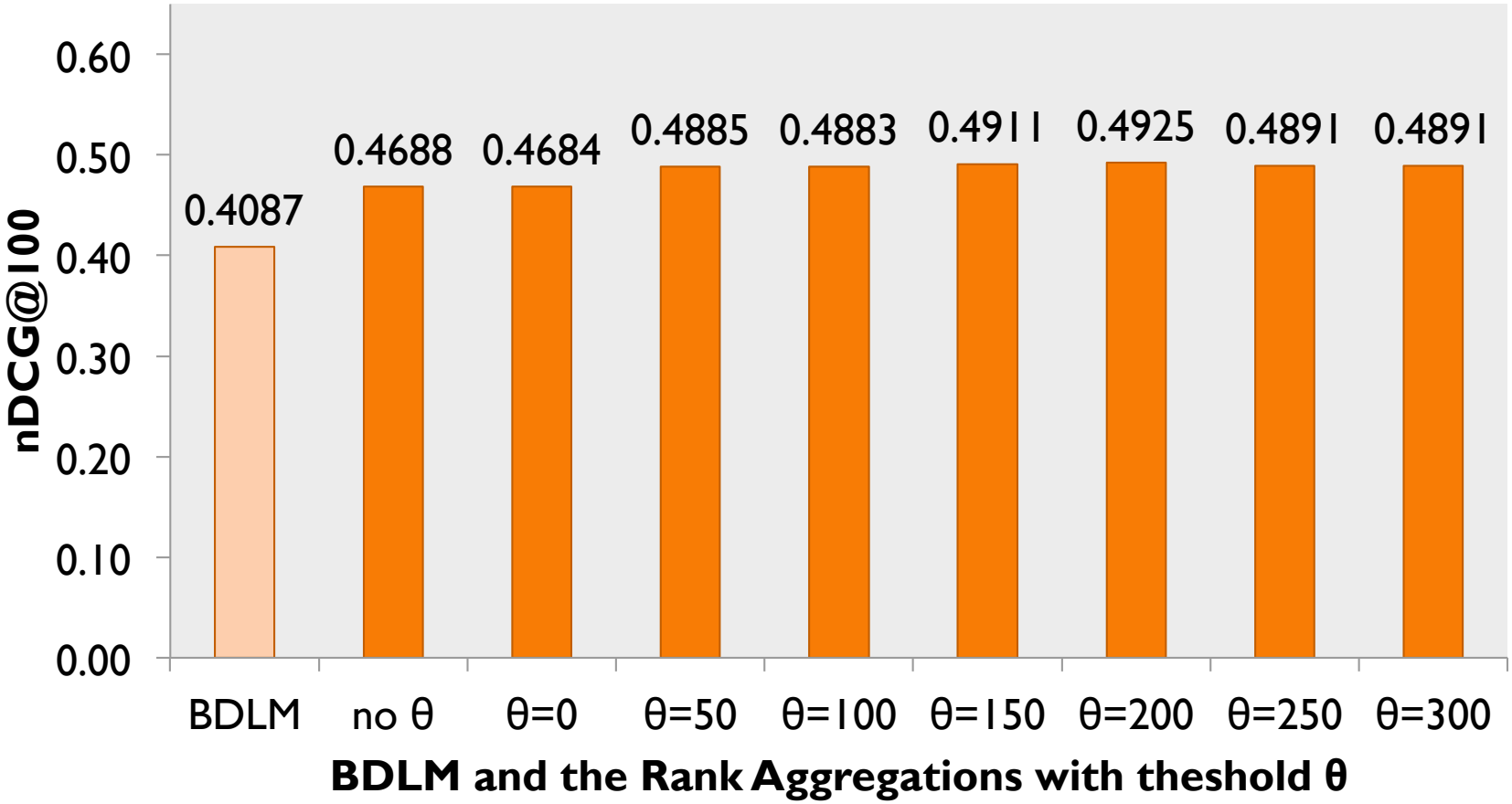
$$t_{ij}^{(k)} = \begin{cases} \dfrac{1}{\#\left\{ j \,\middle|\, j >_{\tau_k} i \ \text{ or } \ i = i \right\}}, \\ \qquad j >_{\tau_k} i \ \text{or} \ j = i \ \text{and z-score}(i) \geq \theta \\ 0, \quad \text{otherwise} \end{cases} \tag{9}$$

- The final score is

$$x = T^T x_0, \quad x_0 = \begin{bmatrix} 1/|D| \\ \vdots \\ 1/|D| \end{bmatrix} \tag{10}$$

Evaluation

Test in NTCIR-8 Corpus



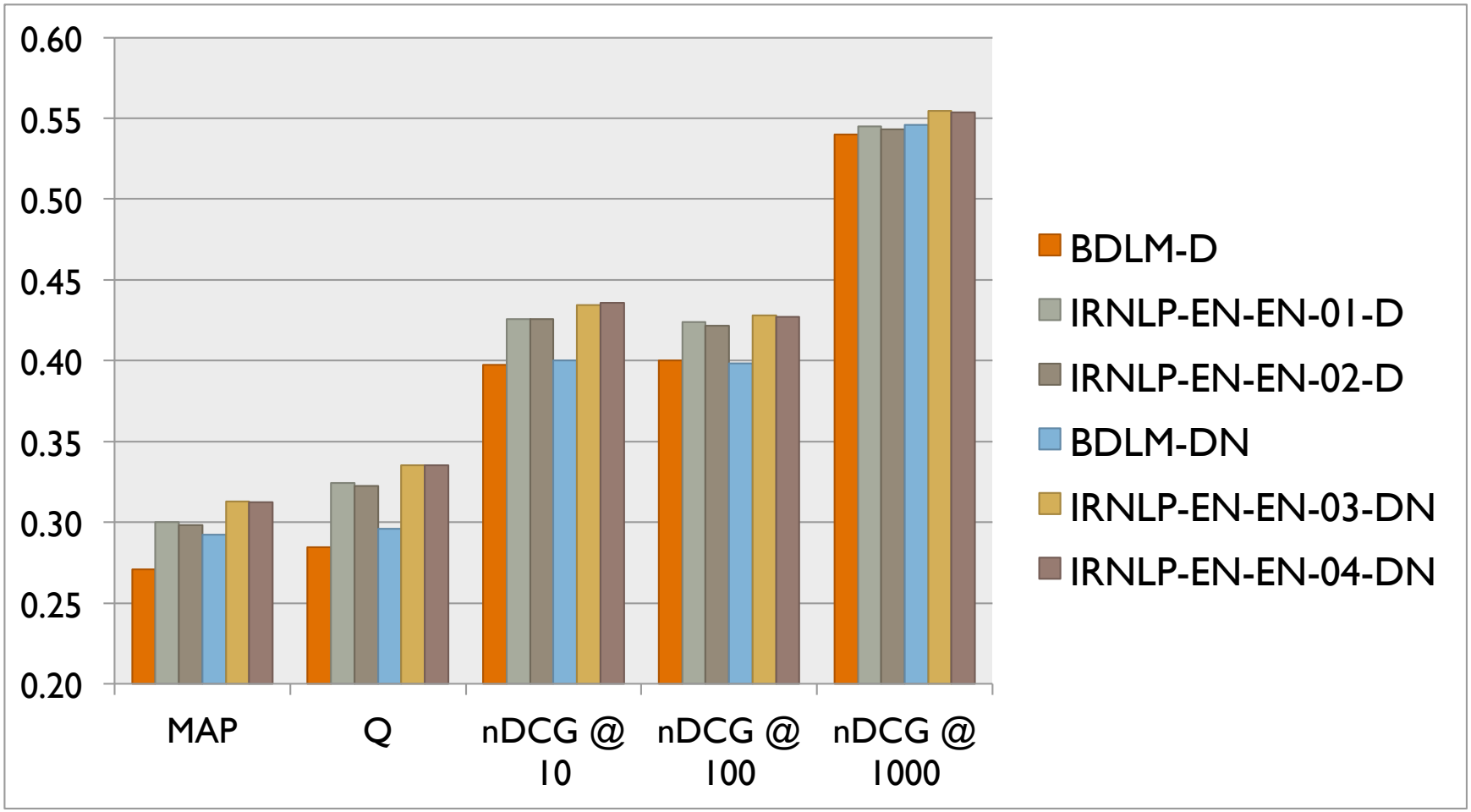
Evaluation

Description of Runs in NTCIR-9

- Summited Runs

RUN	Topic Source	Aggregation	Aggregation Threshold (θ)
IRNLP-EN-EN-1-D	Description only	(BD, RD, BS, & RS) LM	150
IRNLP-EN-EN-2-D	Description only	(BD, RD, BS, & RS) LM	200
IRNLP-EN-EN-3-DN	Description & Narrative	(BD, RD, BS, & RS) LM	200
IRNLP-EN-EN-4-DN	Description & Narrative	(BD, RD, BS, & RS) LM	140
BDLM-D	Description only	BDLM	-
BDLM-DN	Description & Narrative	BDLM	-

Results from NTCIR-9 (1/2)



Evaluation

Results from NTCIR-9 (2/2)

Topics showing high performances

- Those verbs are related to the activities or states of agents clearly (e.g. "*murder*", "*hijack*", "*kill*", and so on).
- The terms are also not ambiguous because they are proper nouns or very specific number of theme (e.g. "4 people" in GeoTime-0033).

Topics showing low performances

- Many errors in the analysis of topics
- The verbs were related to the existence or occurrence of agent or theme. (e.g., "*occur*", "*happen*")
- They sometimes require inference or term expansion.

Conclusion & Future Work

- A new geo-temporal information retrieval method that utilizes **semantic role labeling** and **rank aggregation**.
  - It is useful to analyze documents for semantic roles around main predicates of sentences and generate language models after the analysis.
  - While the SRL-based method is not always superior across different topics, they complement the usual language modeling approach
  - It warrant the proposed rank aggregation method.

- Future Work
  - Term expansion and weighting are necessary.
  - It is also needed to find the optimal weight and thresholds in rank aggregation.

Q&A

Thank you for your attention.